RL of Partially Observable Environments using Spectral Methods

Kamyar Azizzadenesheli

UC. Irvine

Joint work with Prof. Anima Anandkumar and Dr. Alessandro Lazaric.

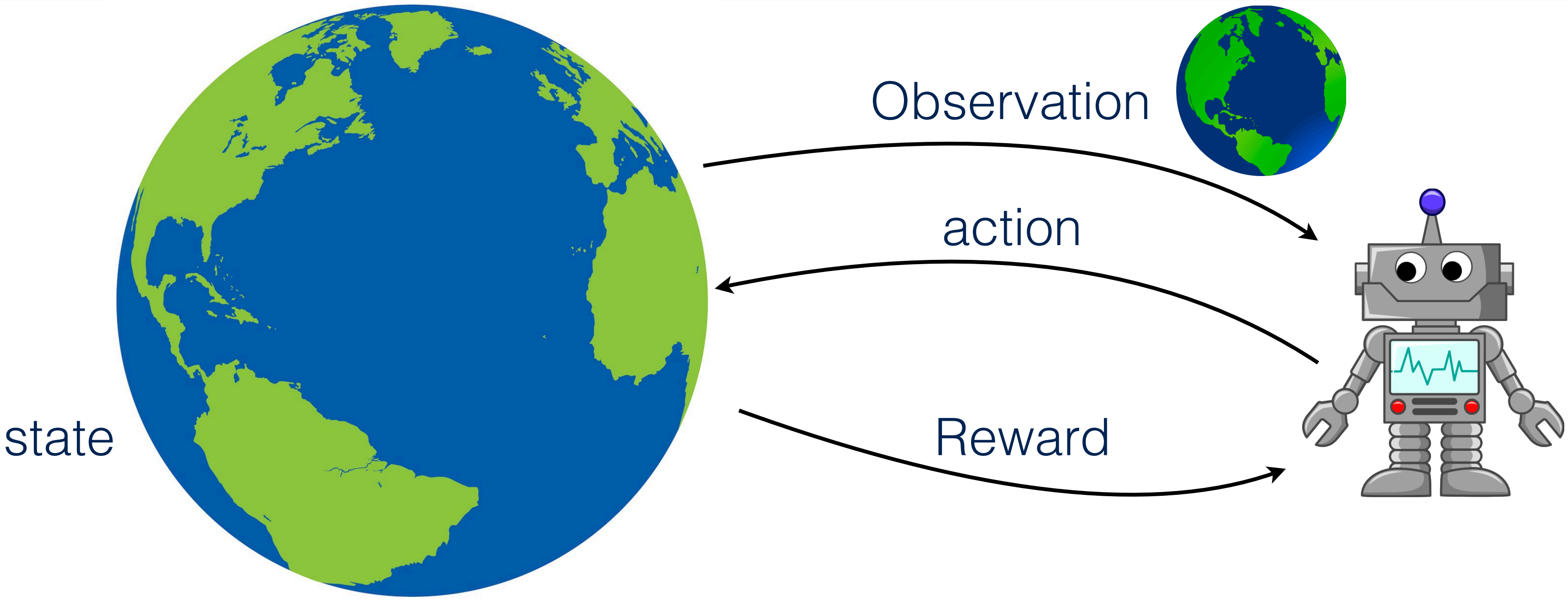
Outline

- Introduction
- Markov Decision Process (MDP)
- Contextual MDP (CMDP)
- Partially Observable MDP (POMDP)

Outline

- **Introduction**
- Markov Decision Process (MDP)
- Contextual MDP (CMDP)
- Partially Observable MDP (POMDP)

Learning in Adaptive Environments



Reinforcement Learning: feedback or reward to reinforce the policy

Goal: maximize the cumulative reward

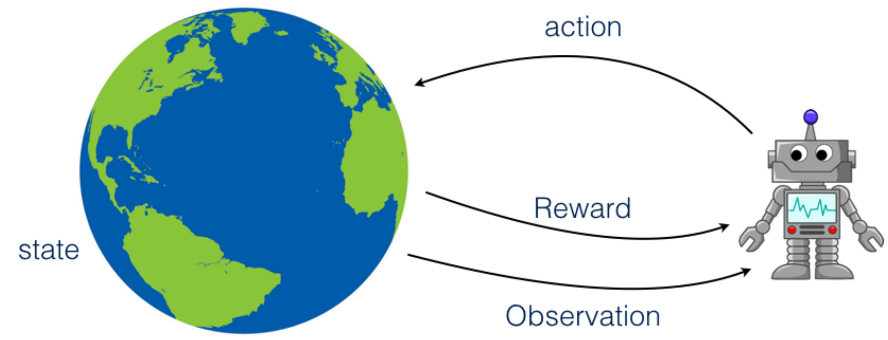
How to evaluate the performance?

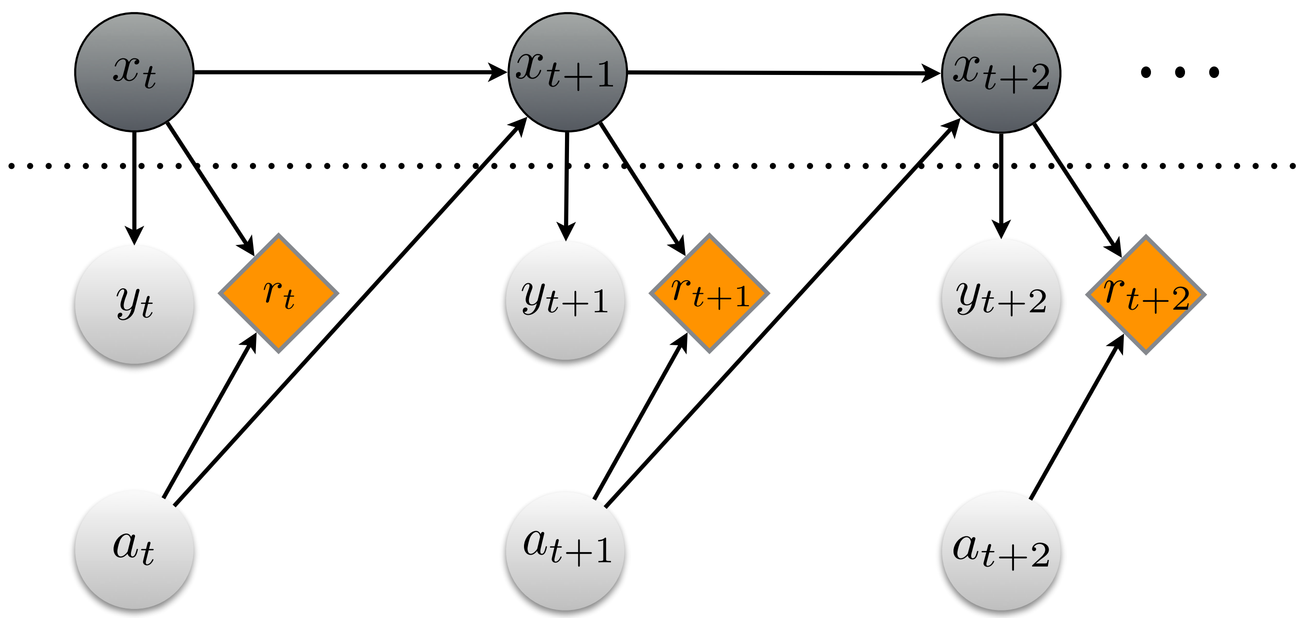
Regret

How much more reward the agent could collect if it knows the environment dynamics

Model-Based RL

Markovian state evolution





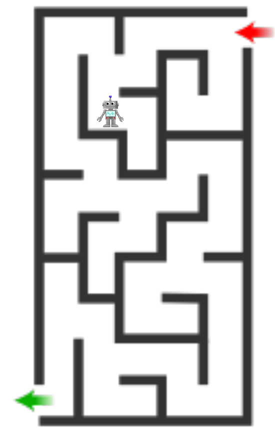
Outline

- **Introduction**
- **Markov Decision Process (MDP)**
- Contextual MDP (CMDP)
- Partially Observable MDP (POMDP)

Fully Observable Model

Playing maze

Observe the map



Playing video games

Access to state of emulator

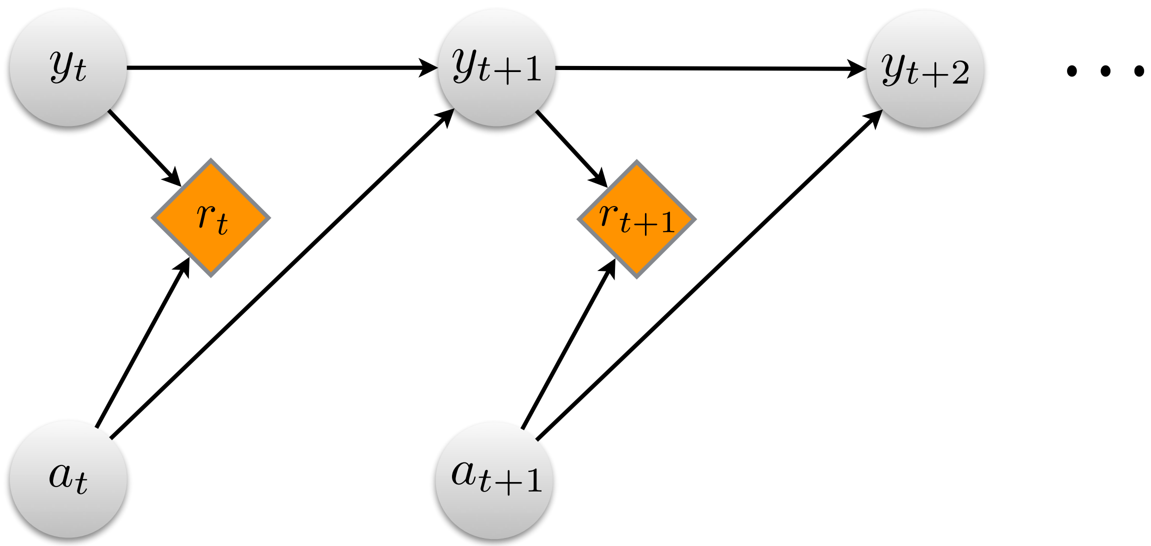
Navigation

Access to the 2D map

Markov Decision Process

Observation is same as the state of the environment

$$y_t = x_t$$



UCRL-MDP

Theoretical results

With high probability the regret of UCRL-MDP is bounded by

$$\mathbf{Reg}_N = \widetilde{\mathcal{O}}\left(DY\sqrt{AN}\right)$$

$$D := \max_{y,y'} \; \min_{\pi} \mathbb{E}[T(y \rightarrow y')|\overline{M}, \pi]$$

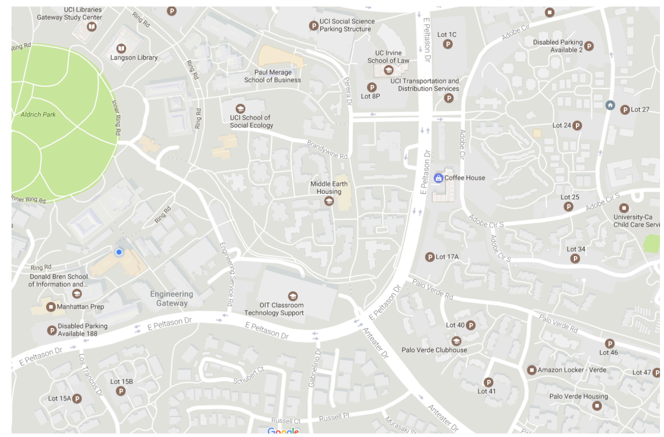
Linear in dimensionality of observation space

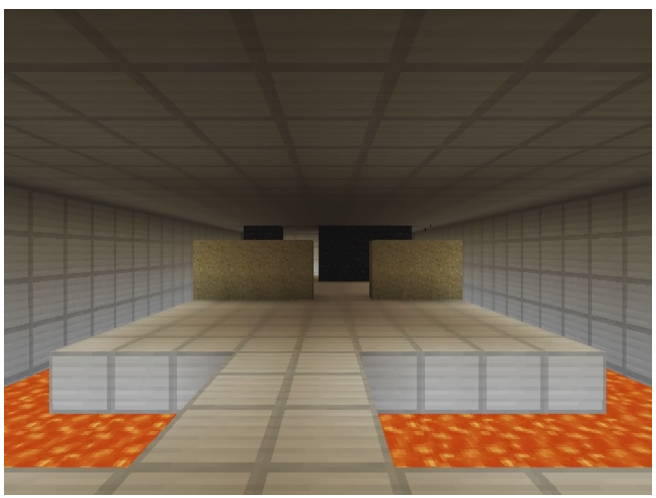
Jaksch, Ortner, Auer, JMLR2010

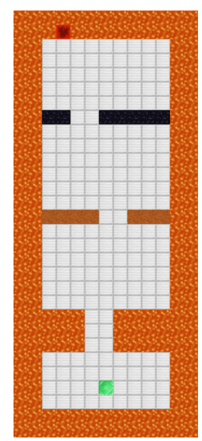
Structured MDPs

Navigation problem

Amazon drone delivery, Grid world



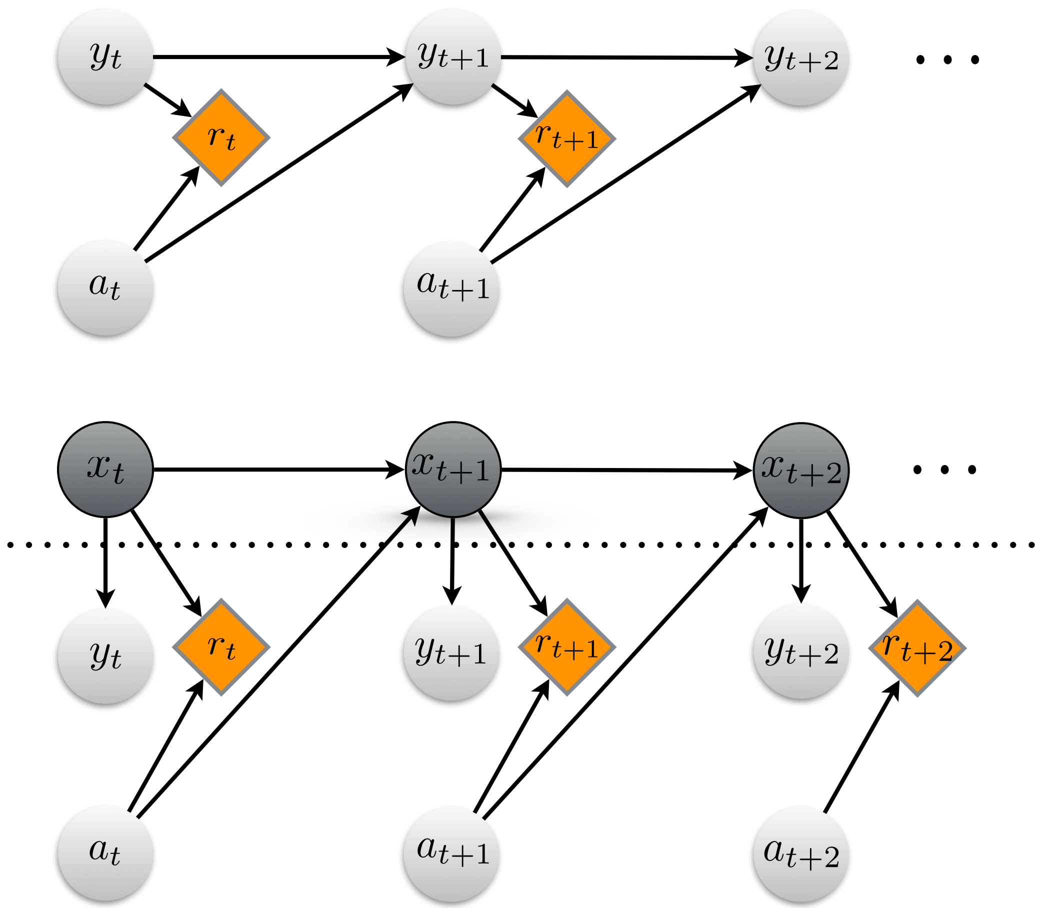


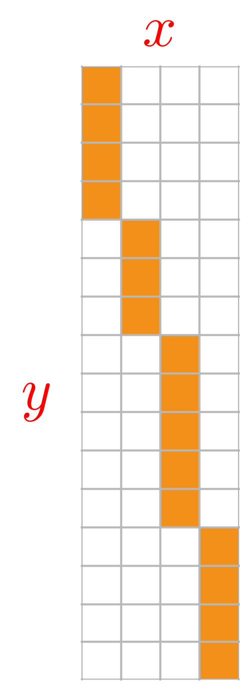


[Johnsonet al., 2016]

Contextual-MDP

Underlying small MDP





Contextual-MDP

Theoretical results

With high probability the regret of UCAgg is bounded by

$$\mathbf{Reg}_N = \widetilde{\mathcal{O}}\left(DC{\color{red}Y}\sqrt{AN}\right)$$

$$D := \max_{y,y'} \min_{\pi} \mathbb{E}[T(y \rightarrow y')|\overline{M}, \pi]$$

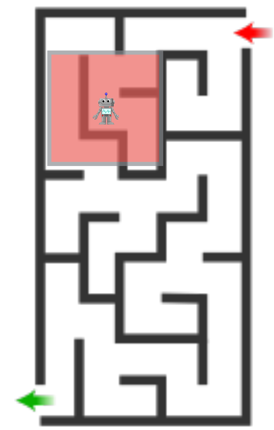
Better computational complexity

Ortner, Sprinter 2013

Partially Observable MDP

Playing maze

Observe part of the map



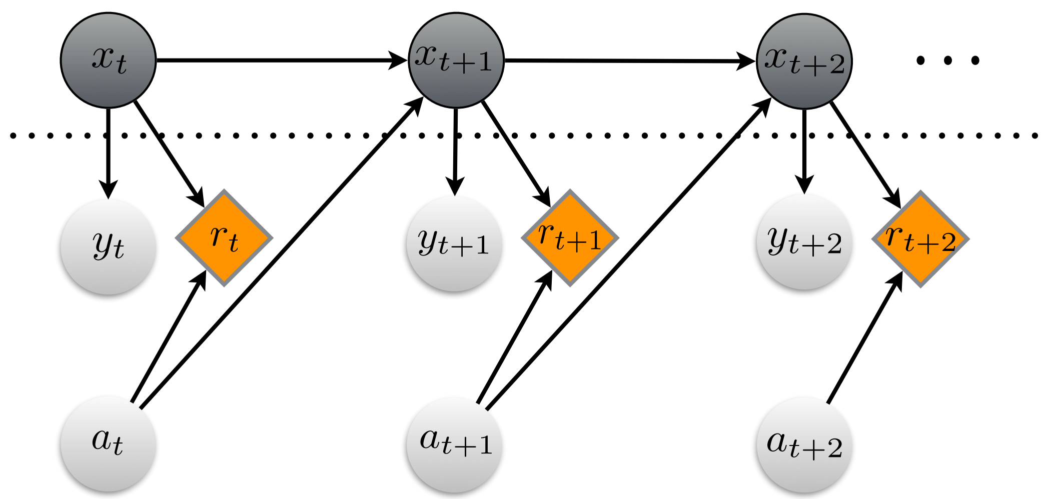
Playing video games

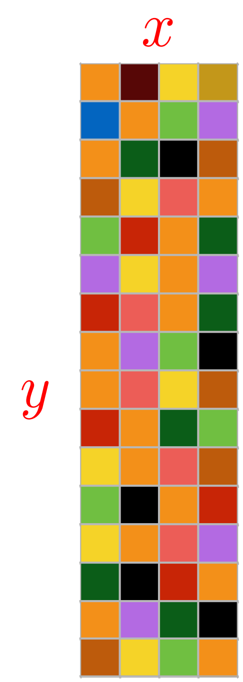
Observe the screen

Self driving car

Sensory observation

Partially Observable MDP

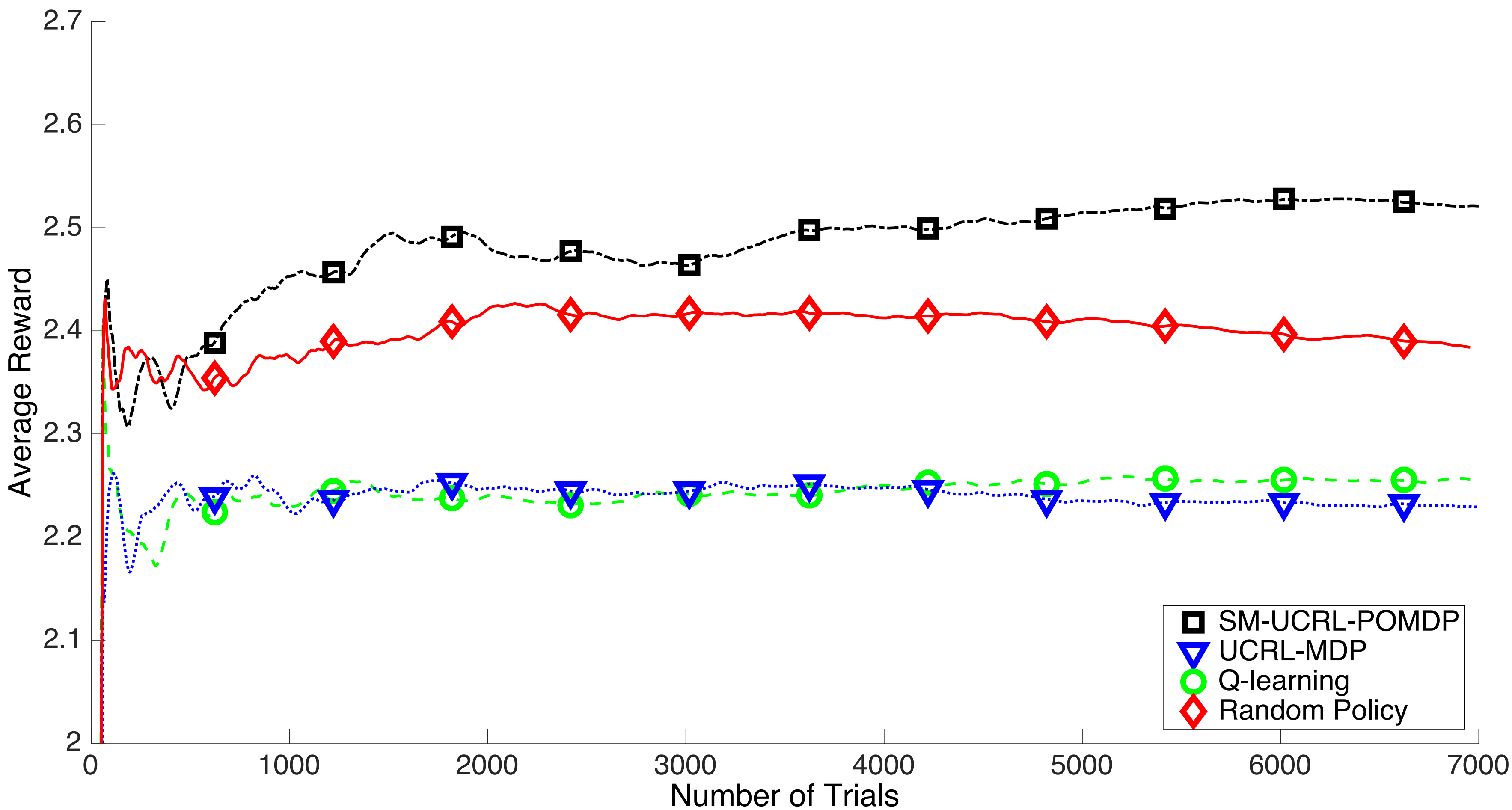




POMDP

Models

X = 2 , Y = 4 , A = 2

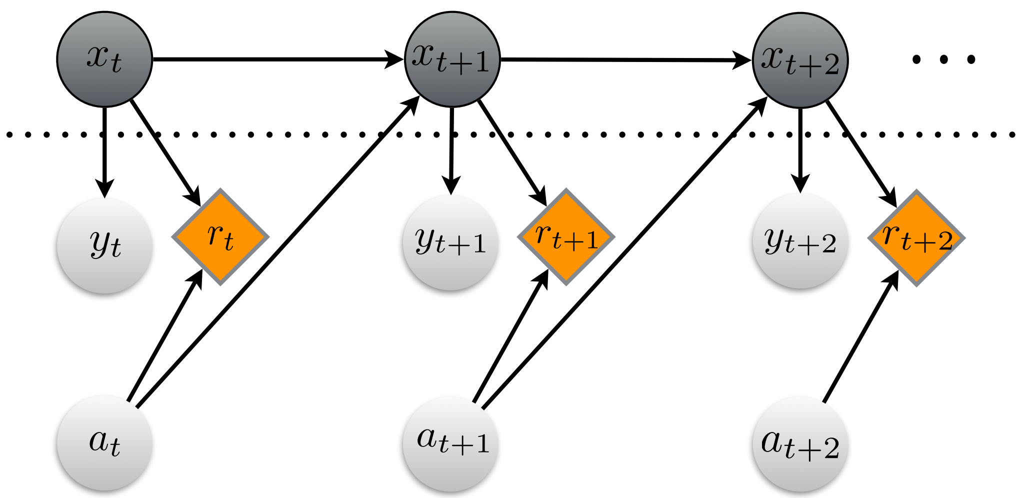


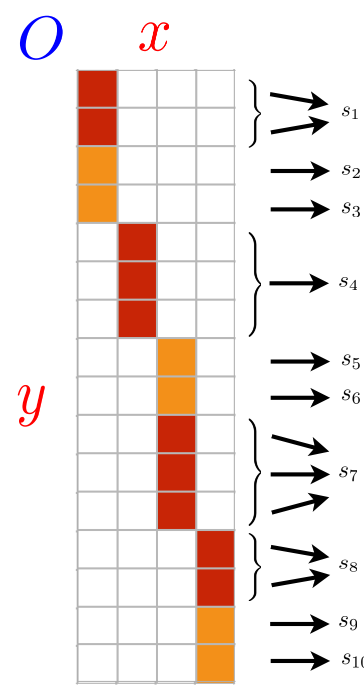
Outline

- Introduction
- Markov Decision Process (MDP)
- **Contextual MDP (CMDP)**
- Partially Observable MDP (POMDP)

Learning the Mapping

The agent's desire is to learn the "context-to-state" mapping

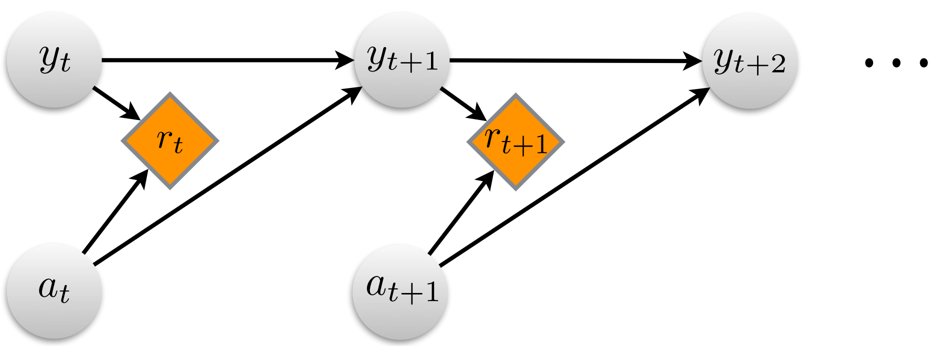




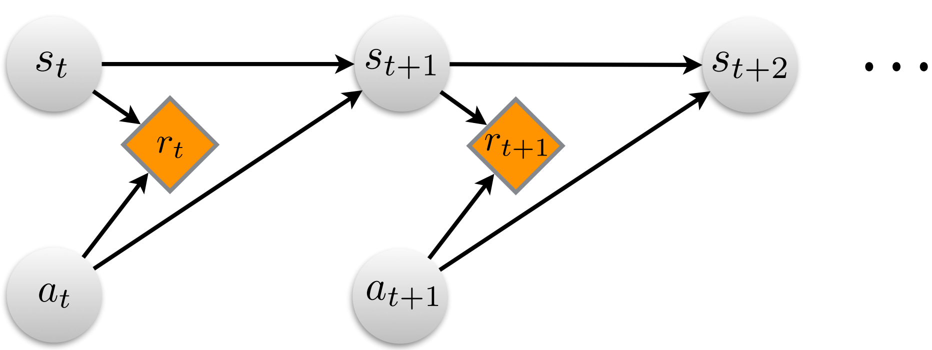
Learning the Mapping

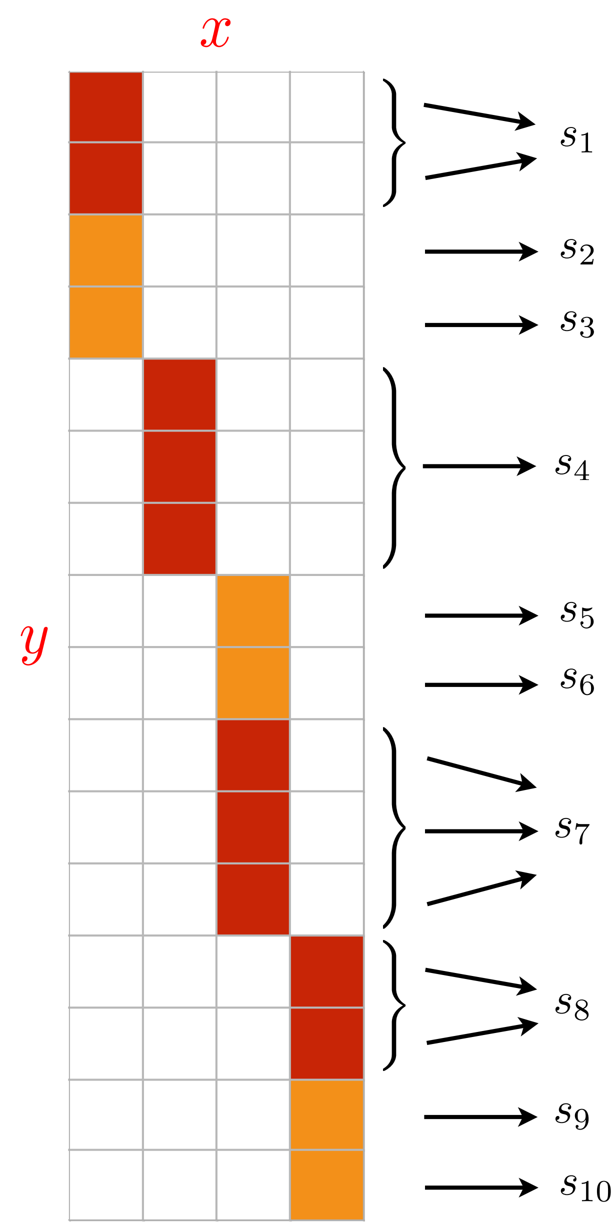
Apply an initial policy

Cluster the contexts



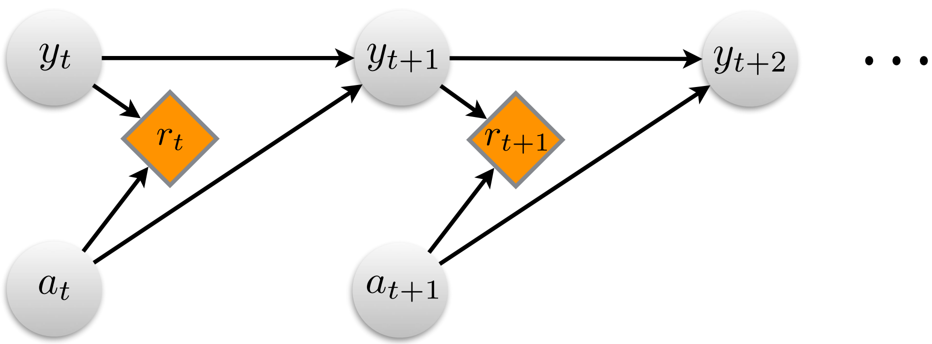
Auxiliary MDP



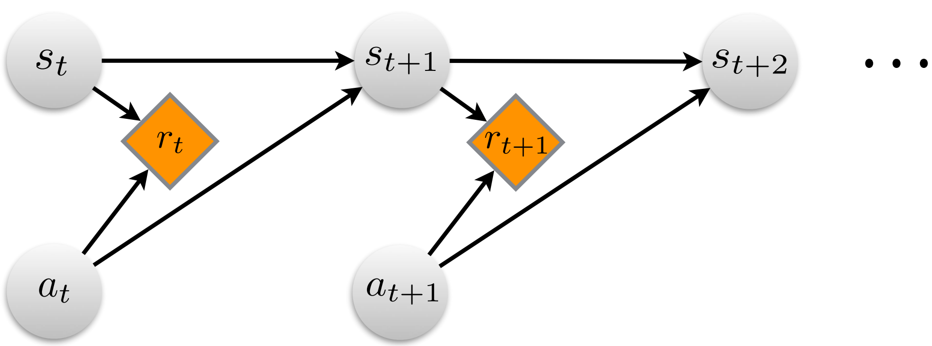


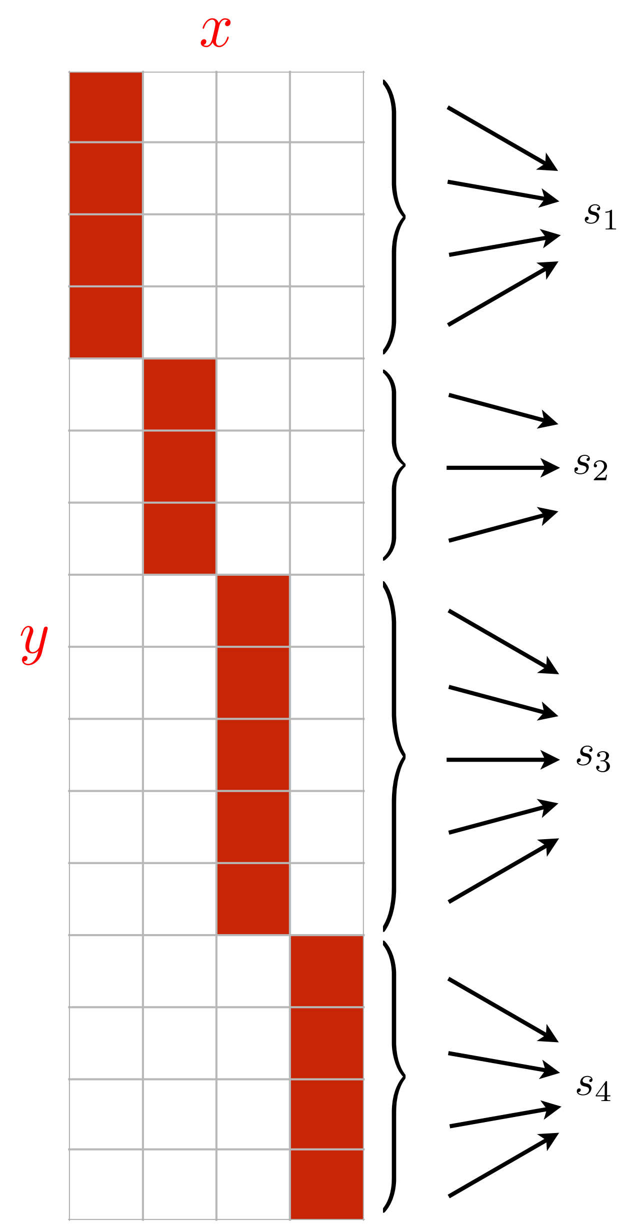
Learning the Mapping

Cluster the contexts



Auxiliary MDP ⬇ True small MDP





Theoretical Results

With high probability, the regret of SM-UCRL-CMDP is bounded by

$$\mathbf{Reg}_N = \widetilde{\mathcal{O}}\left(D_C X \sqrt{AN}\right)$$

$$D_C := \max_{x,x'} \min_{\pi} \mathbb{E}[T(x \rightarrow x')|\overline{M}, \pi]$$

Compared to UCAgg

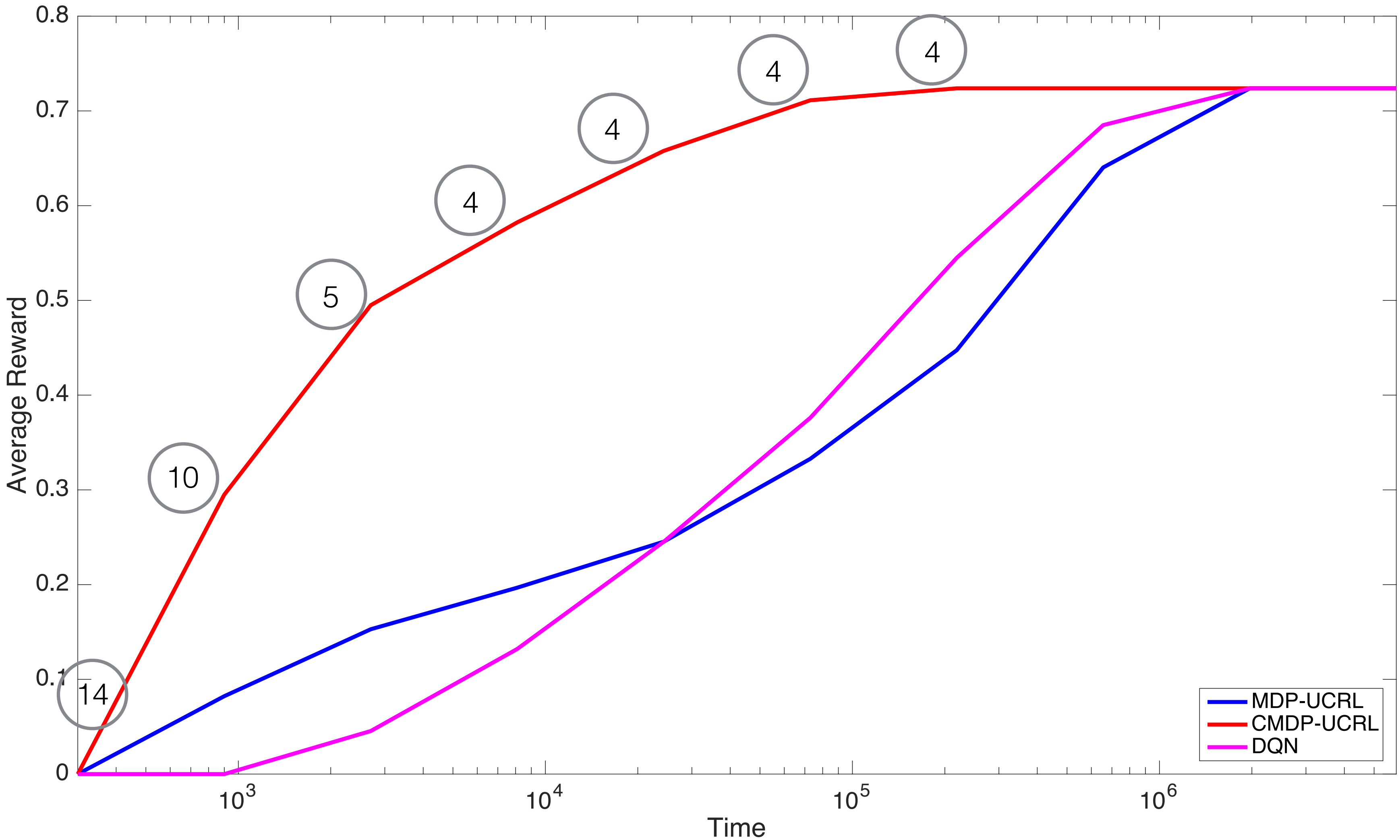
$$\mathbf{Reg}_N = \widetilde{\mathcal{O}}\left(DCY \sqrt{AN}\right)$$

$$D := \max_{y,y'} \min_{\pi} \mathbb{E}[T(y \rightarrow y')|\overline{M}, \pi]$$

K. Azizzadenesheli, A. Lazaric, A. Anandkumar, 2016

Experimental Results

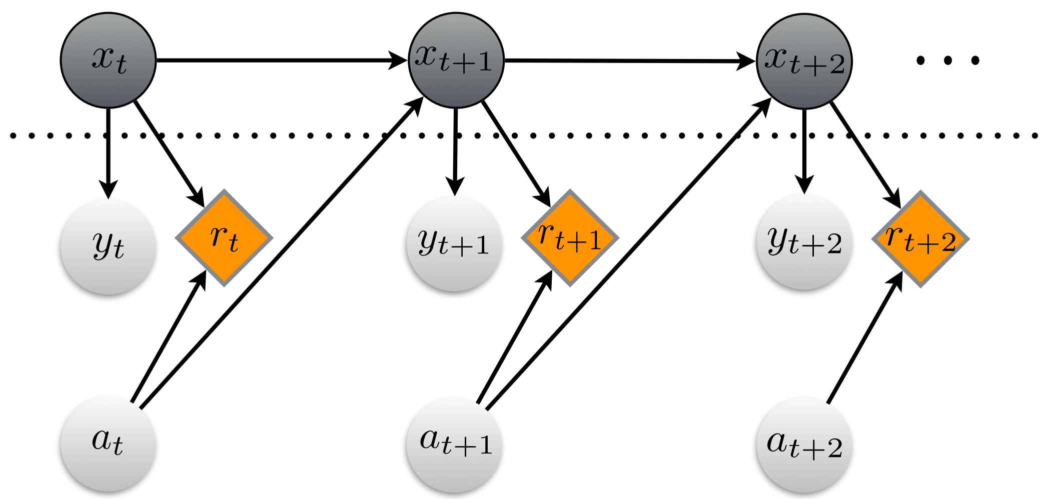
Synthetic Env. X = 4, Y =20 , A = 4

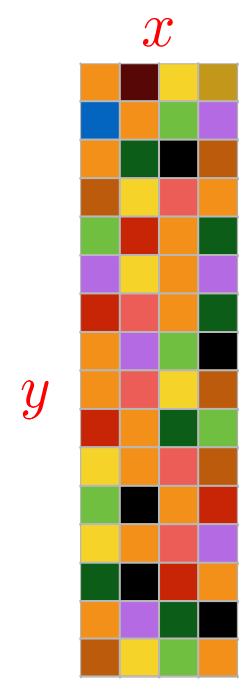


Outline

- Introduction
- Markov Decision Process (MDP)
- Contextual MDP (CMDP)
- **Partially Observable MDP (POMDP)**

Partially Observable MDP

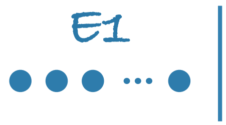




POMDP

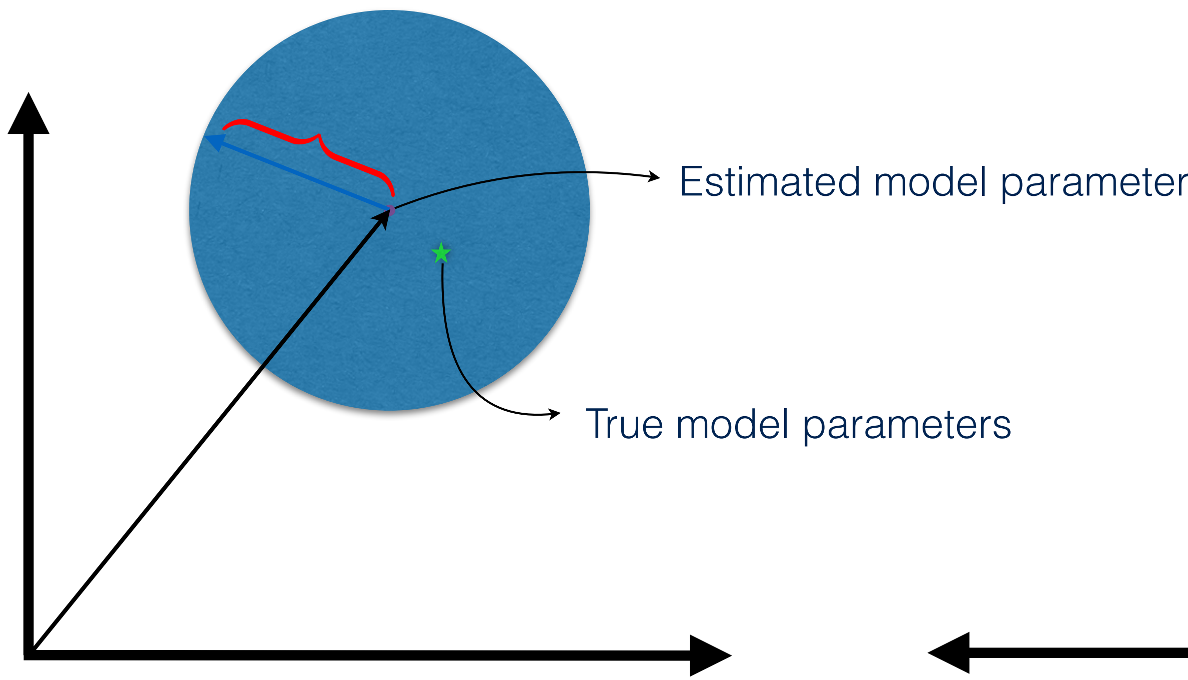
Exploration-Exploitation

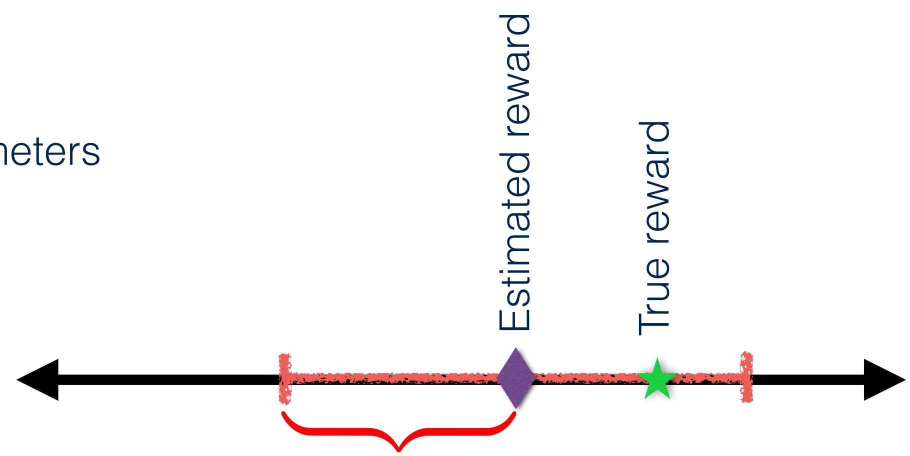
Epoch-based RL



Apply an initial policy and collect samples

High probability confidence



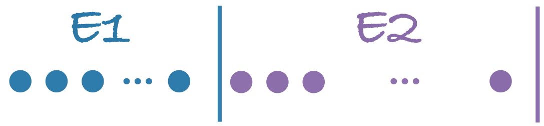


Construct a set of plausible models

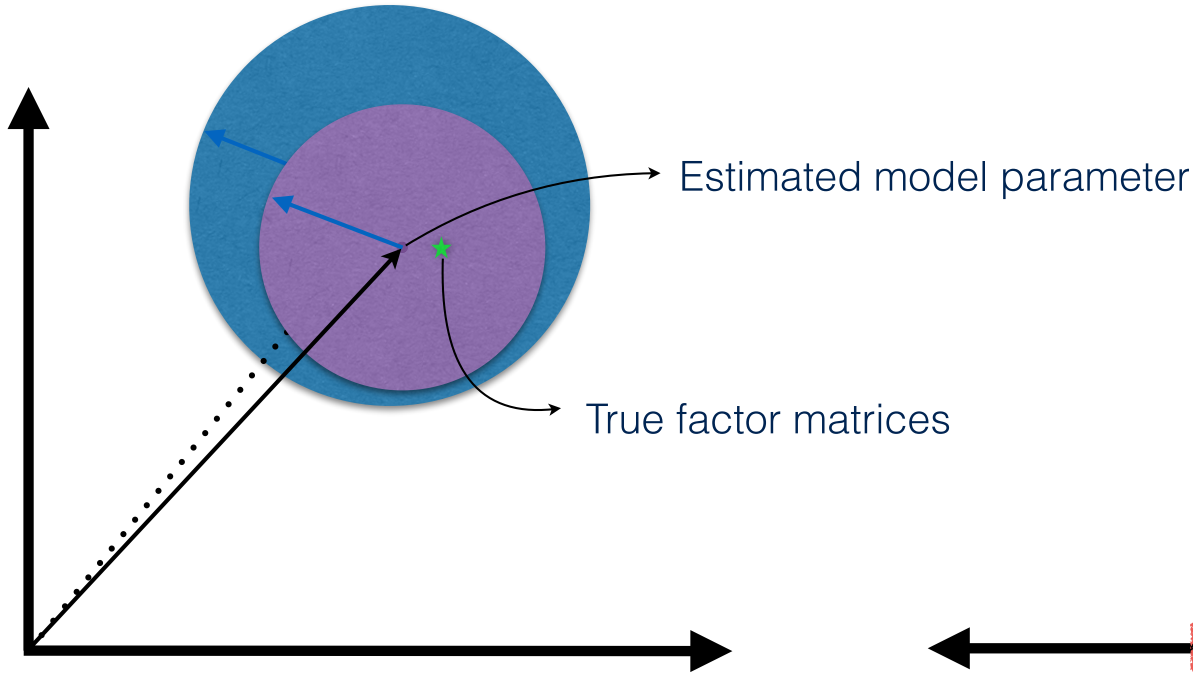
Optimism in face of uncertainty (OFU)

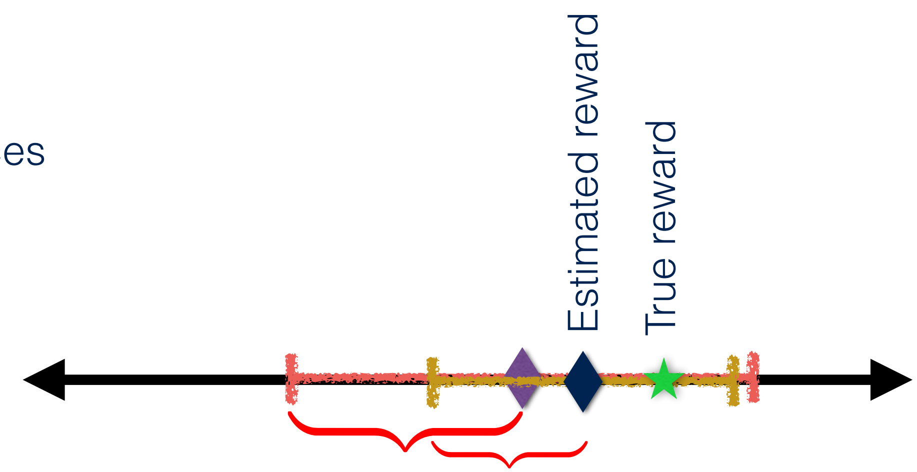
Exploration-Exploitation

Epoch-based RL



Apply a policy and collect samples



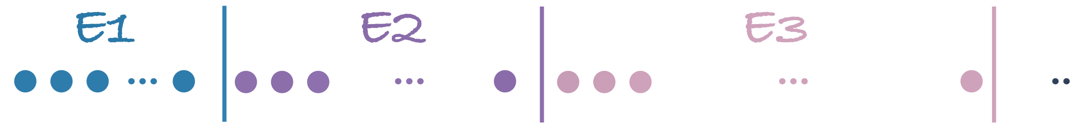


Construct a set of plausible models

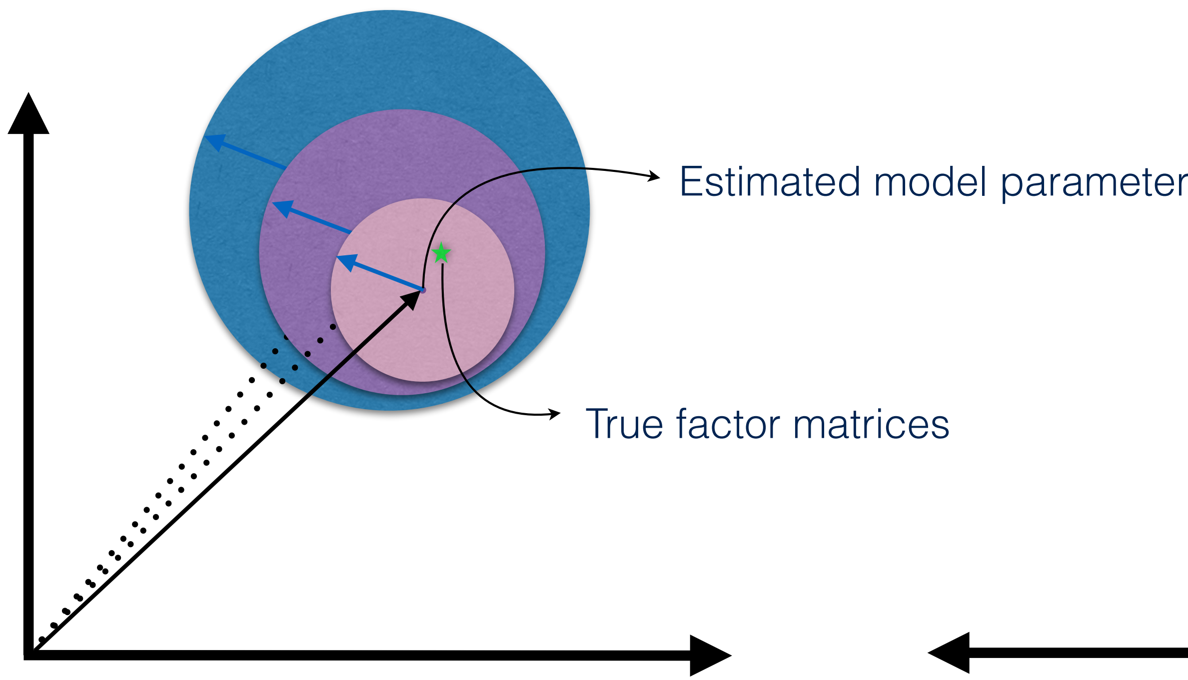
Optimism in face of uncertainty (OFU)

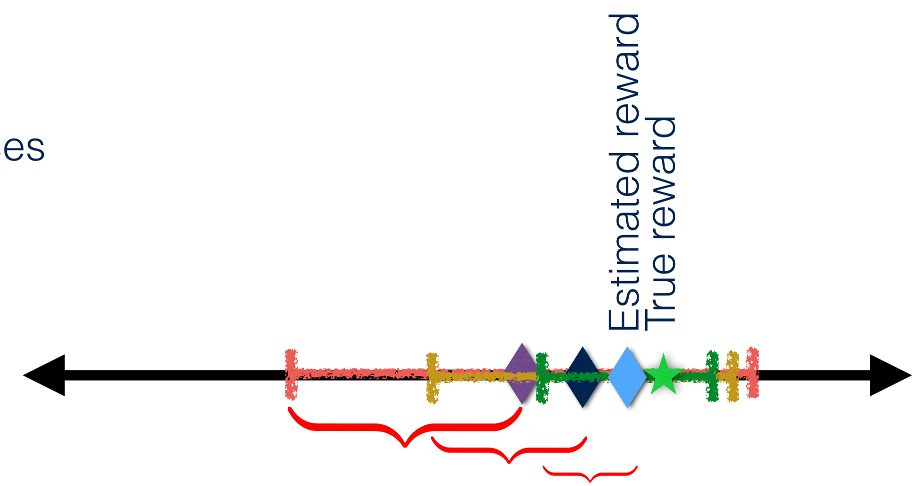
Exploration-Exploitation

Epoch-based RL



Apply a policy and collect samples





Construct a set of plausible models

Optimism in face of uncertainty (OFU)

Theoretical Results

With high probability, the regret of SM-UCRL-POMDP is bounded

$$\mathbf{Reg}_N = \widetilde{\mathcal{O}}\left(D_P X \sqrt{ANXY}\right)$$

$$D_P := \max_{(x,a),(x',a')} \min_{\pi} \mathbb{E}[T((x,a) \rightarrow (x',a'))|\overline{M},\pi]$$

Compared to MDP

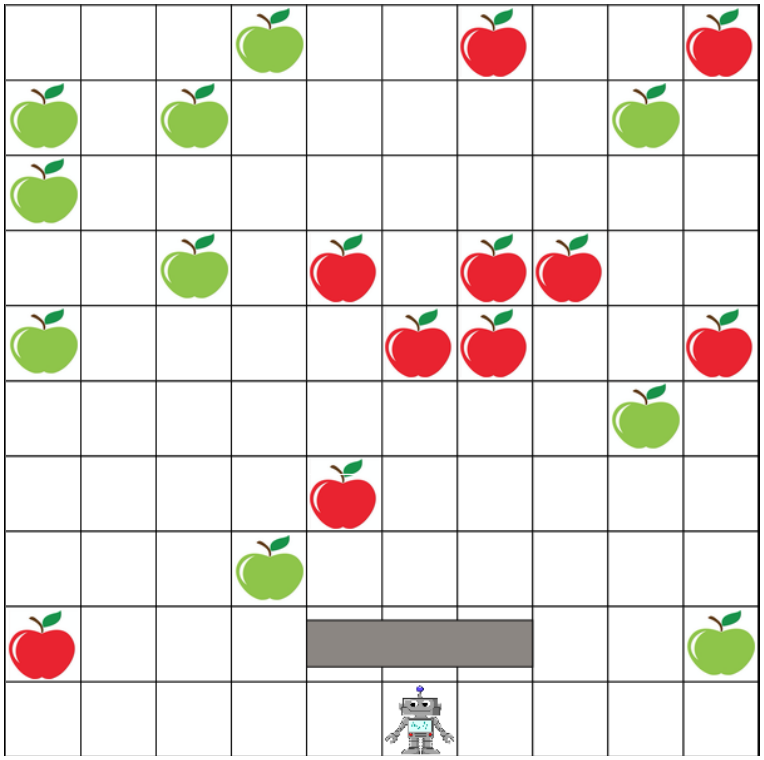
$$\mathbf{Reg}_N = \widetilde{\mathcal{O}}\left(DY\sqrt{AN}\right)$$

$$D := \max_{y,y'} \min_{\pi} \mathbb{E}[T(y \rightarrow y')|\overline{M},\pi]$$

K. Azizzadenesheli, A. Lazaric, A. Anandkumar, COLT 2016

Empirical Results

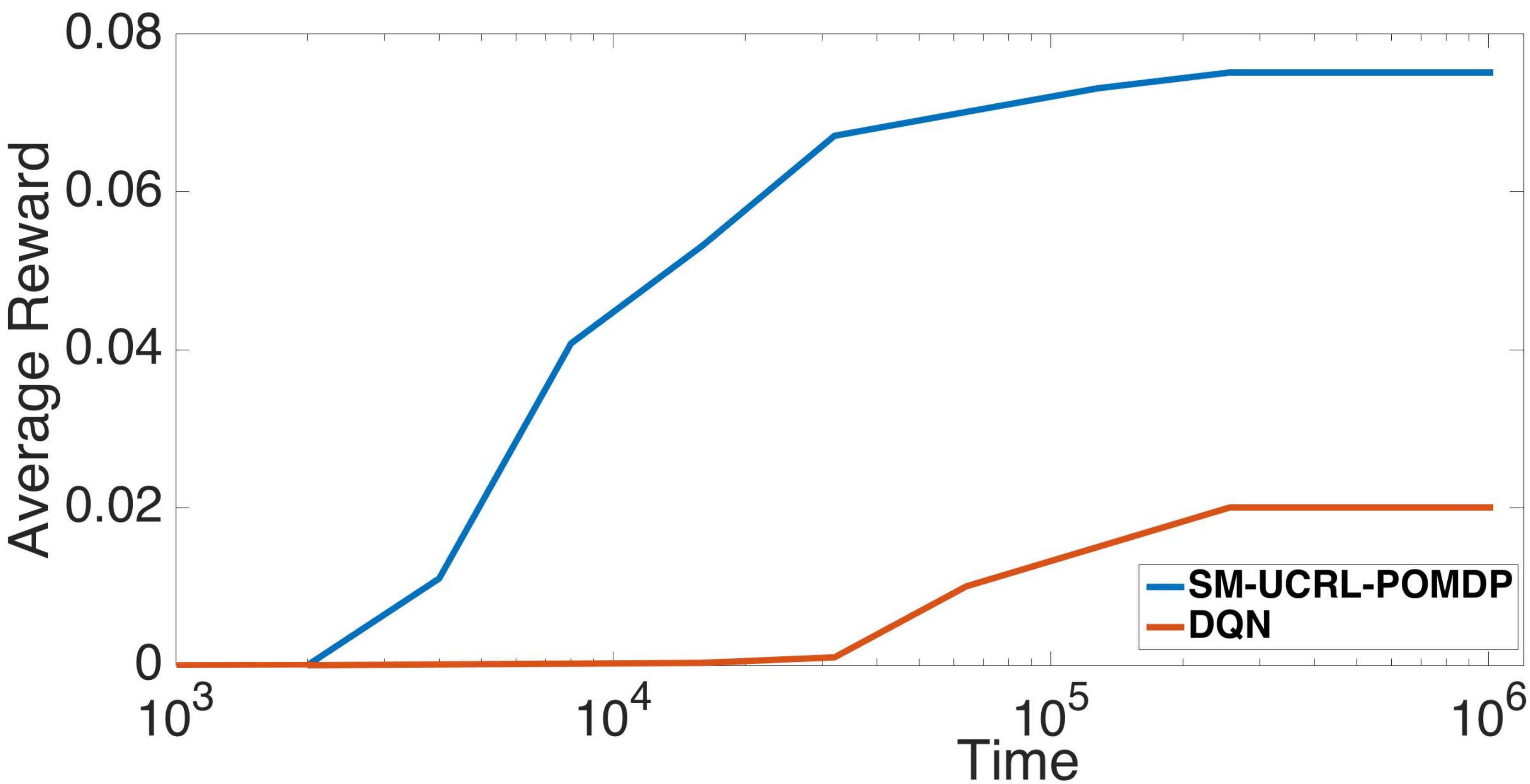
- Triple boxes observation



K. Azizzadenesheli, A. Lazaric, A. Anandkumar, NIPS DeepRL, 2016

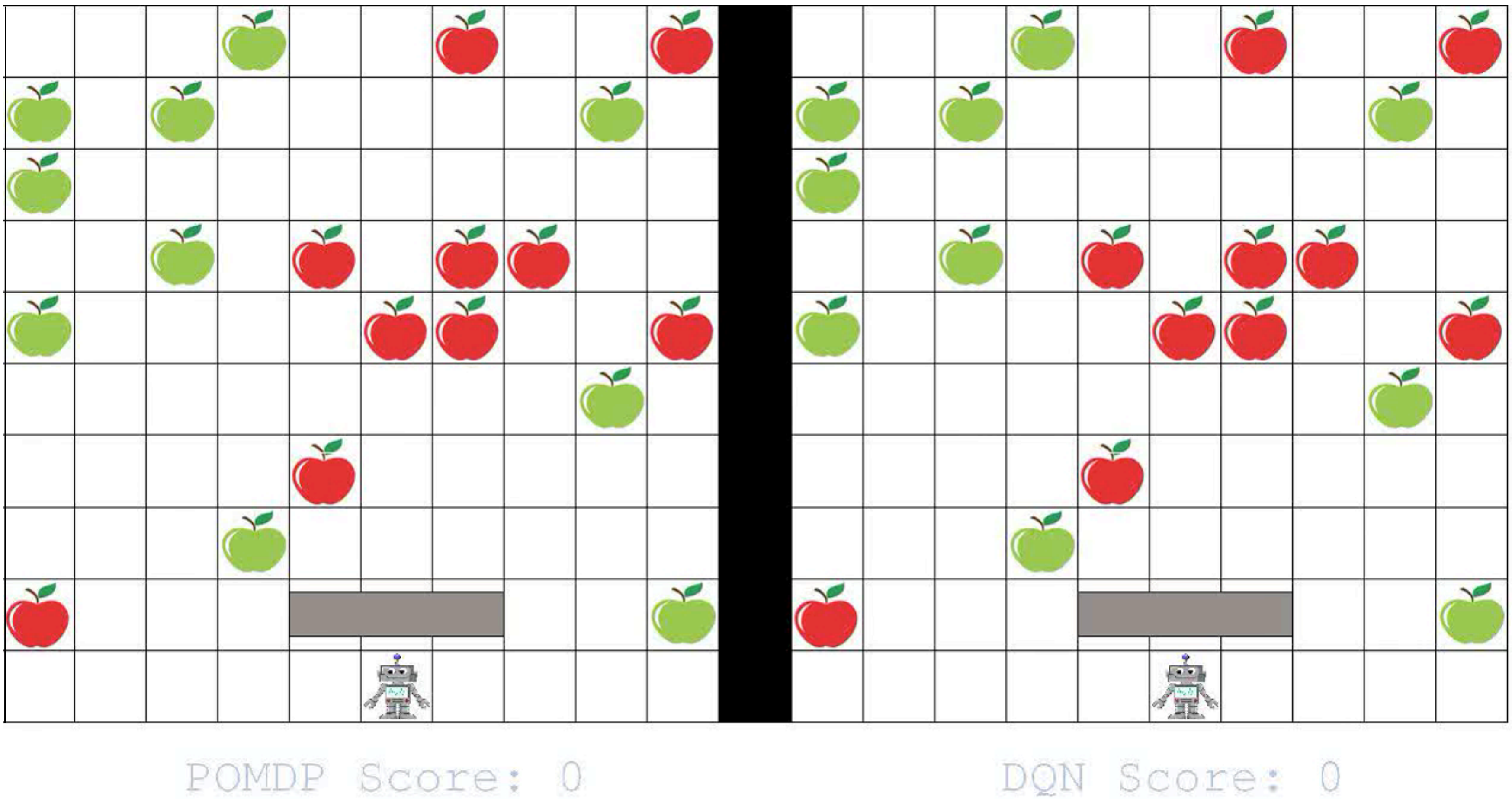
Empirical Results

- 8 hidden states SM-POMDP-UCRL
- 30x30x30 DQN (RMSprop w. hyperbolic tangent units)



K. Azizzadenesheli, A. Lazaric, A. Anandkumar, NIPS DeepRL, 2016

Empirical results



K. Azizzadenesheli, A. Lazaric, A. Anandkumar, NIPS DeepRL, 2016

Thank you